INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA Project - Analysis of AMCAT Data

About Me

Coming from a B.Sc. Agriculture background, I've recently completed a rigorous data science course at AlmaBetter. This program equipped me with essential skills in Python, machine learning, and data analysis, fostering my transition into the dynamic field of data science. Now, I'm eager to apply my diverse background and newfound expertise to drive insights and innovation in data-driven environments.

My decision to pursue a career in data science stems from a profound interest in the field's ability to unlock insights and drive innovation across various sectors. Data science offers a dynamic landscape where I can apply advanced analytical techniques to tackle complex challenges and make data-driven decisions. I am drawn to the interdisciplinary nature of data science, which combines elements of mathematics, statistics, computer science, and domain expertise to extract actionable insights from large datasets. By immersing myself in the world of data science, I aim to contribute to cutting-edge projects, drive business growth, and solve real-world problems through the power of data analytics and machine learning algorithms.

Currently, I am interning at Innomatics Research Labs, where I am gaining practical experience in data science methodologies and techniques. This hands-on opportunity allows me to further enhance my skills and apply theoretical knowledge to real-world projects in a professional setting.

INNOMATICS RESEARCH LABS

OBJECTIVE OF THE PROJECT

This analysis aims to gain insights and understanding from the provided dataset, particularly focusing on the relationship between various features and the target variable, which is Salary. Specifically, the goals of this analysis include:

- Describing the dataset and its features comprehensively.
- Identifying any patterns or trends present in the data.
- Exploring the relationships between independent and target variables (Salary).
- Identifying any outliers or anomalies in the data.

SUMMARY OF DATA

The Aspiring Mind Employment Outcome 2015 (AMEO) dataset, released by Aspiring Minds, focuses on employment outcomes for engineering graduates. It includes dependent variables such as Salary, Job Titles, and Job Locations, along with standardized scores in cognitive skills, technical skills, and personality skills. With around 40 independent variables and 4000 data points, these variables encompass both continuous and categorical data. The dataset also includes demographic features and unique identifiers for each candidate.

Exploratory Data Analysis

DATA CLEANING AND PREPROCESSING

Datatype Conversion

1. To enhance the data integrity and ensure consistency, we replaced all instances of the values 0 and -1 with null values throughout the entire dataset.
2. To ensure the accuracy and consistency of our analysis, we converted the data types of the 'Date of Joining' (DOJ) , 'Date of Leaving' (DOL) and DOB fields from their original format to datetime objects.
3. We conducted a data cleaning process specifically targeting the "jobcity" column to address any spelling mistakes present. This involved label encoding to standardize spellings and ensure consistency throughout the dataset.

Feature Engineering

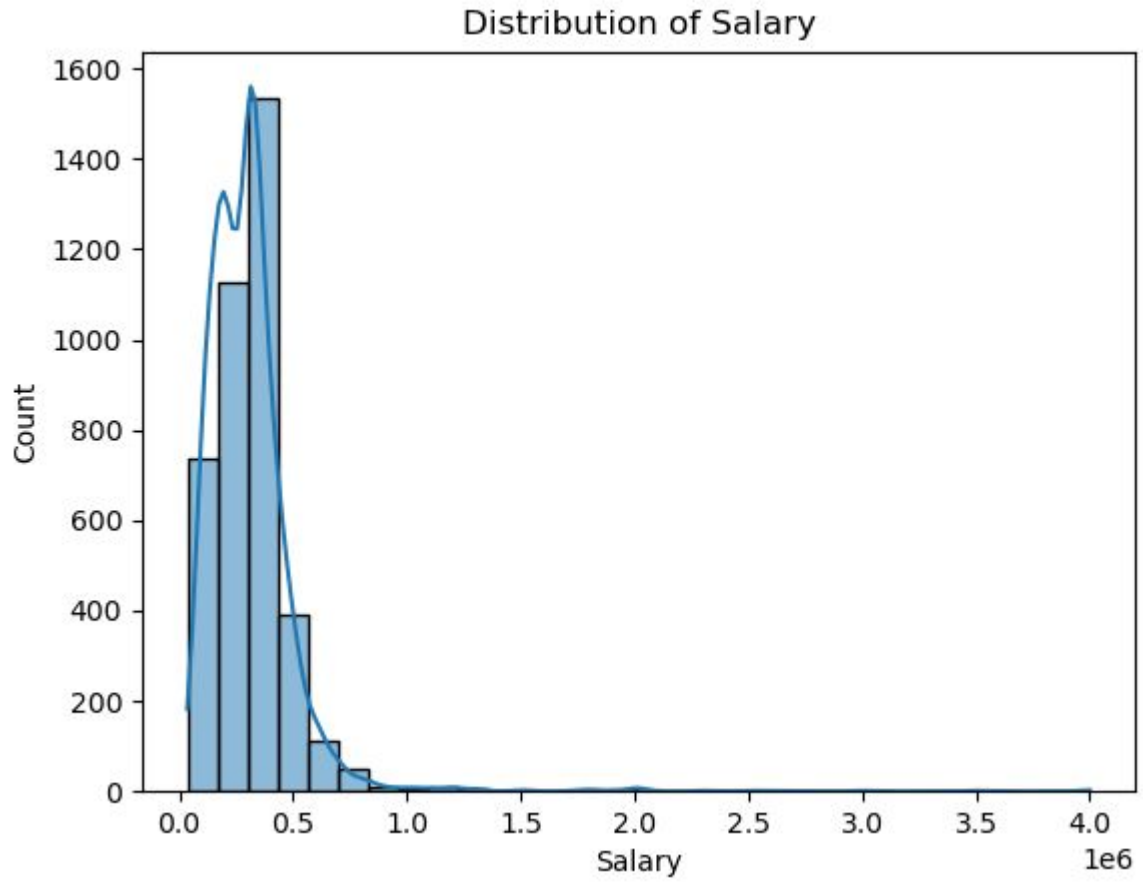
Age Calculation:

An additional column representing age has been incorporated into the dataset by subtracting the year of birth (DOB) from current date.

Univariate Analysis

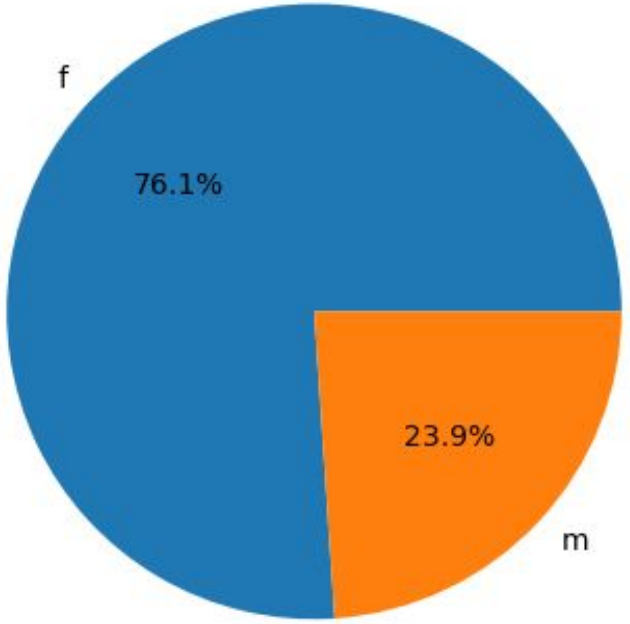
1. Salary

- The distribution shows positive skewness, with a longer right tail.
- There are a few outliers visible in the histogram.
- Most of the data is between 0.0 and 0.5(i.e majority of people have salary between 35,000 to 50,000)
- The minimum salary observed is 35,000 and maximum is 40,00,000.



2. Age

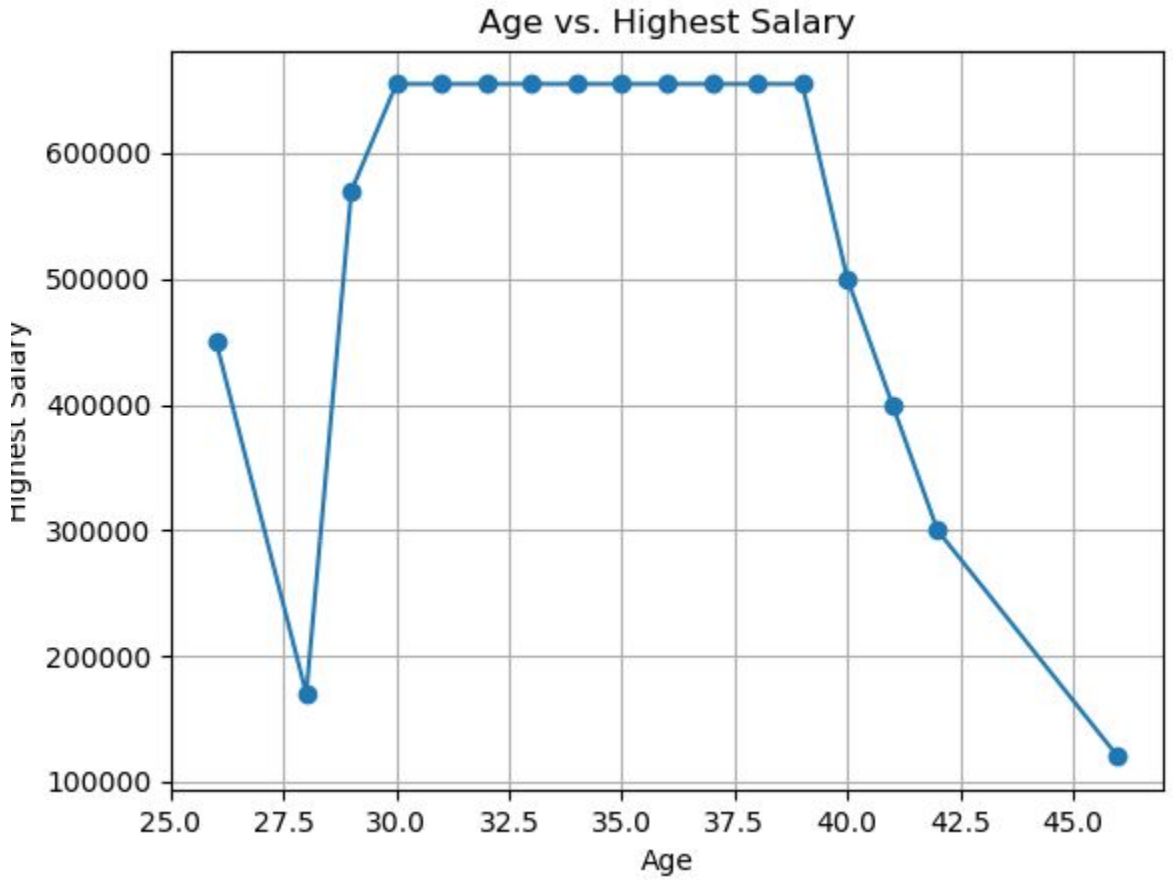
- The distribution indicates an imbalanced gender distribution where one gender significantly outweighs the other in terms of representation.
- Such skewed gender distributions can impact the analysis and interpretation of the data, potentially leading to biases or incomplete insights.
- The distribution also indicates a higher representation of males compared to females in the workforce.



Bivariate Analysis

1. Age & Salary

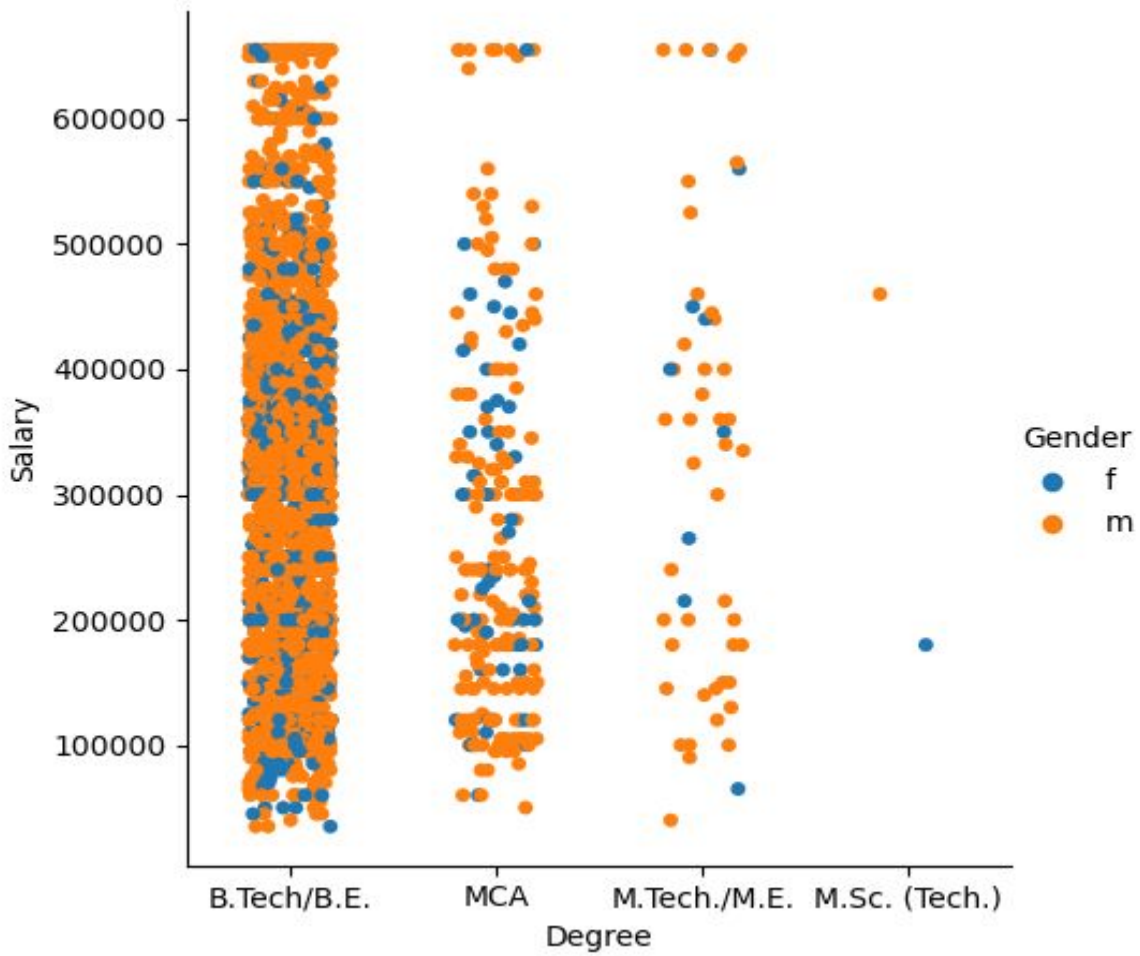
- The distribution shows that highest salary making age group is between 30 to 40, which indicates that the experienced age group is getting paid more compared to the age group below 30 which can be freshers or less experienced.



INNOMATICS RESEARCH LABS

2. Salary, Gender & Degree

In the scatter plot representing salary against degree, with different genders depicted by distinct colors, a notable trend emerges in the distribution of data points across different educational levels. Upon closer examination, it becomes evident that the concentration of data points is noticeably higher within the bachelor's degree category compared to other categories.



Conclusion

- The salaries in the dataset vary widely, ranging from 80,000 to 1,100,000 Indian Rupees annually.
- The dataset contains both male and female individuals, with a majority being male.
- Employees are situated in various cities across India, including Bangalore, Gurgaon, Hyderabad, Noida, Chennai, Pune, Indore, Kolkata, and others.
- There's a diversity of job titles, ranging from software engineers and systems engineers to quality assurance engineers, project engineers, and more.
- Employees' duration at their respective jobs ranges from recent hires to individuals who have been with their companies for several years.
- Most individuals hold Bachelor's degrees in Engineering (B.Tech/B.E.), with specializations in fields like computer engineering, electronics and communication engineering, mechanical engineering, and information technology. There's also a mention of some individuals holding a Master's in Computer Applications (MCA).
- Academic performance varies among individuals, with percentages ranging from 52.33% to 95.8% in the 10th and from 43.42% to 95.2% in the 12th grade.
- While some individuals have a specific degree in computer science or related fields, others have pursued more specialized fields like mechanical engineering, electrical engineering, etc.
- The dataset also includes scores in various skills such as English, Logical, Quantitative, Domain Knowledge, Computer Programming, Electronics and Semiconductors, Computer Science, Mechanical Engineering, Electrical Engineering, Telecom Engineering, Civil Engineering, and others.
- There are also personality trait scores provided in the dataset, including conscientiousness, agreeableness, extraversion, neuroticism, and openness to experience.

THANK YOU

INNOMATICS RESEARCH LABS